Clustering under constraints for multivariate and heterogeneous time series - Application to hydrological data

Thesis subject – ICube – SDC team

1 Context

Data classification/clustering is one of the most widely used types of unsupervised learning to help analyze a new dataset. It involves partitioning the set of analyzed objects into groups called clusters. There are a huge number of clustering algorithms, each with its own advantages and disadvantages [4, 9]. For example, the well-known K-Means algorithm is very easy to understand, implement, and use, but remains highly sensitive to its initial parameters (number of clusters, choice of seeds, etc.) and presupposes the existence of convex clusters in the data space. This last point clearly shows that the space into which the data are projected, notably through the choice of similarity measure, has a strong influence on the clustering result.

In particular, the clustering of sequences raises problems related to the measurement of similarity between two individuals. For example, in river monitoring, certain phenomena have an annual frequency linked to the natural water cycle, but can be shifted in time due to geographical distance and local meteorology. Similarity measurements must be able to take these potential shifts or slight distortions in time into account. Numerous methods have been proposed in the literature to take these particularities into account [7], e.g. Dynamic Time Warpping (DTW) [13], Longest Common SubSequence (LCSS) [1], or more recently representations by shapelets [11] or neural networks [8].

However, these methods offer little or no solution to the problem of missing values. In the case of time series, the absence of certain values can occur at different time steps between the different individuals studied, creating even more artificial time shifts. In the literature, data interpolation is generally used to regenerate missing values [10]. Nevertheless, data generation poses a number of problems. First, interpolated data makes little sense when the system may be subject to very large and/or frequent variations in values. Second, this is exacerbated when studying multivariate time series, as some variables may have missing values at different time steps for the same individual. Thus, partial data generation can create inconsistencies between the different measured variables. In the literature, proposals have been made to take into account the time spread without data generation, but only in supervised mode, the main solutions being

the addition of timestamps [3] or the use of masking [2] in time series. However, these approaches cannot be used as such in unsupervised mode, as they have a very negligible impact on the data space. To date, and to the best of our knowledge, no method has been proposed in this case.

2 Subject

The main objective of the thesis will be to develop new approaches for measuring the similarity between two multivariate time series, taking into account missing values distributed heterogeneously in time and between variables, in an unsupervised/weakly supervised context. This work will focus on defining solutions for integrating temporal information (spacing between two time steps, temporal frequencies, etc.) into the calculation of similarity. We will also be looking at how to integrate the expert's knowledge though annotations, also known as constraints (e.g. proximity/distance between two individuals based on external information), concerning both temporal and spatial links between different individuals. Thus, it aims to improve the correspondence between the clustering obtained and the expert's expectations.

As part of this thesis, we will be experimenting with river monitoring data. These data raise a number of difficulties, due to their number, diversity and heterogeneity, both spatially and temporally. In practical terms, a monitoring network was set up following the Water Framework Directive (WFD2000). At each station of the network, various samples are taken regularly (at most every 2 months), but at different times : physicochemical parameters (nitrates, phosphate, pH, etc.), microparameters (pesticides, for example), biological parameters (concerning aquatic flora and fauna). Other stations are monitored by national hydrological services, where flow rates and water levels are measured on a daily basis. These data have been exploited by various mining methods, but require discretization [5, 12]. A first clustering approach on digital temporal data was studied in [6] and opened interesting avenues for hydroecology.

In this work, we will focus in particular on the following questions :

— clustering of time series of parameters vectors, given that these parameters are not always measured at the same time
— taking into account temporal (seasons) or geographical (hydrographic regions) constraints
— coupling physico-chemical data with biological or hydrological data (different temporalities)
— explore the limits of sequence number and size

3 Supervision and location

The thesis is directed by Florence Le Ber, HDR, and supervised by Baptiste Lafabrègue. It will benefit from the expertise of Corinne Grac (hydroecologist, UMR LIVE) for the application part.

The thesis will be carried out at the ICube laboratory in Illkirch (near Strasbourg).

4 Candidate profile

Required profile :
- Master 2 in Computer Science
- Training in data science, data mining, machine learning

Training and required skills :
- Excellent knowledge of machine learning and knowledge modeling
- Excellent programming skills in Python or R
- Excellent communication and writing skills in English (French not mandatory)
- An interest in the application's subject

5 Application procedure

Applications should be submitted by email to both :
- florence.le-ber@unistra.fr
- lafabregue@unistra.fr

They must include : – A Curriculum Vitae ; – List of 2 or 3 references to contact (position, email address) ; – Transcripts of graduate studies ; – Link to MSc thesis, and publications if applicable ; – Link to personal software repositories (e.g. GitHub)
The thesis is subject to an evaluation to be held on 11th and 12th of June 2024, and requires a good academic CV. The complete application must be submitted by May 23rd.
See the doctoral school website : `http://ed.math-spi.unistra.fr/candidature/sujets-des-contrats-doctoraux-de-recherche/`

Références

[1] Lasse Bergroth, Harri Hakonen, and Timo Raita. A survey of longest common subsequence algorithms. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 39–48. IEEE, 2000.

[2] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1) :6085, 2018.

[3] Mamadou Ben Hamidou Cissoko, Vincent Castelain, and Nicolas Lachiche. Multi-way adaptive time aware lstm for irregularly collected sequential icu data. *Available at SSRN 4567952*, 2023.

[4] Xu Dongkuan and Tian Yingjie. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2 :165–193, 2015.

[5] Mickaël Fabrègue, Agnès Braud, Sandra Bringay, Corinne Grac, Florence Le Ber, Danielle Levet, and Maguelonne Teisseire. Discriminant temporal patterns for linking physico-chemistry and biology in hydro-ecosystem assessment. *Ecological Informatics*, 24 :210–221, 2014.

[6] Corinne Grac, Agnès Braud, Pierre Gançarski, Agnès Herrmann, and Florence Le Ber. Comparing the physico-chemistry dynamics of running waters (North-East of France) based on sequence clustering. *Ecological Informatics*, 72 :101921, Dec 2022.

[7] Christopher Holder, Matthew Middlehurst, and Anthony Bagnall. A review and evaluation of elastic distance functions for time series clustering. *Knowledge and Information Systems*, pages 1–45, 2023.

[8] Baptiste Lafabregue, Jonathan Weber, Pierre Gançarski, and Germain Forestier. End-to-end deep representation learning for time series clustering : a comparative study. *Data Mining and Knowledge Discovery*, 36(1) :29–81, 2022.

[9] Thomas Lampert, Thi-Bich-Hanh Dao, Baptiste Lafabregue, Nicolas Serrette, Germain Forestier, Bruno Crémilleux, Christel Vrain, and Pierre Gançarski. Constrained distance based clustering for time-series : a comparative and experimental study. *Data Mining and Knowledge Discovery*, 32 :1663–1707, 2018.

[10] Mathieu Lepot, Jean-Baptiste Aubin, and François HLR Clemens. Interpolation in time series : An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, 9(10) :796, 2017.

[11] Arnaud Lods, Simon Malinowski, Romain Tavenard, and Laurent Amsaleg. Learning dtw-preserving shapelets. In *Advances in Intelligent Data Analysis XVI : 16th International Symposium, IDA 2017, London, UK, October 26–28, 2017, Proceedings 16*, pages 198–209. Springer, 2017.

[12] Cristina Nica, Agnès Braud, Xavier Dolques, Marianne Huchard, and Florence Le Ber. Exploring Temporal Data Using Relational Concept Analysis : An Application to Hydroecology. In *CLA : Concept Lattices and their Applications*, volume 1624 of *CEUR Workshop Proceedings*, pages 299–311, Moscow, Russia, 2016.

[13] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1) :43–49, 1978.